9

COMPUTER-PRODUCED DISTRIBUTION MAPS OF DISEASE

Howard C. Hopps

Reprinted from
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 161, Article 2, Pages 779-799
September 30, 1969

22

COMPUTER-PRODUCED DISTRIBUTION MAPS OF DISEASE*

Howard C. Hopps

Division of Geographic Pathology
The Armed Forces Institute of Pathology
Washington, D. C.

I am a pathologist. I know relatively little about computer technology, but I do know enough to realize that it is the means to solve one of my major problems—a problem in medical intelligence—if I'm smart enough to specify realistic objectives and to see to it that the input data is adequate to the task. My associates and I have worked a little over two years now in developing a system to solve our problem. We've come a long way, but there's a long way yet to go.

A great deal of blood, sweat, tears—and money—has been poured into the problem of pattern recognition. Our problem is a rather different one: the rapid production of patterns. These patterns would be derived from virtually an infinite assortment of patternable data available in the abundant literature dealing with the ecology of disease. We are not so naive as to propose that an infinite number of map patterns will be produced: rather we propose that, of the enormous number of potential patterns (assuming an adequate data base), the user will request those relatively few patterns which may meet his particular needs.

We have chosen maps as the principal pattern form to display information for two reasons: First, because the diseases we are mainly interested in are strongly influenced by a wide variety of ecological factors, e.g., temperature, rainfall, humidity, the amount and mineral content of surface water, agricultural practices, population densities of various plants and animals including man, the kinds of people involved (not only age and sex, but race, ethnic group, and tribe)—and a hundred other factors closely tied to geographic location. Second, because mapmakers have a unique advantage over those persons who communicate information by other forms of graphic display for three general reasons:

1. Extensive and continual usage of map forms, beginning in early childhood, has conditioned most "educated" people to an intuitive understanding of maps.

*Supported by the Advanced Research Project Agency through a contract to Universities Associated for Research and Education in Pathology (DA 49-092-ARO-130) and monitored by the U. S. Army Research Office.

2. The map is ideally suited to a consideration of multiple factors simultaneously (e.g., place—both geographic and political—in relation to topography, population density, the location of towns and cities, the location and character of transportation routes, and time zones).

3. Through the use of rather simple devices, such as isarithms (more properly termed isopleths), one can achieve a three dimensional effect (quantity being the third dimension, quality and location the other two) in a two dimensional presentation.

We believe that a mechanism (system) which can produce many kinds of map-patterns quickly, in response to specific query, will offer two very important advantages: First, such a mechanism will make it possible to have current information about the distribution of specific diseases and the distribution of known important causally related agents or conditions (e.g., in the case of infectious disease, insect vectors or animal reservoirs). Second, the rapid availability of a large number and wide variety of disease/environmental maps will give the observer an opportunity to compare location-patterns of unknown but possibly related ecological factors and, in this way, suggest causal relationships that might otherwise never have come to mind.

From a broader point of view, the MOD (Mapping Of Disease) project is an effort to illuminate the *geographic pathology* of disease. Geographic pathology is, in a sense, a kind of comparative pathology—one in which *place* (rather than species) is the primary variable. Geographic pathology attempts to answer the questions: What (disease); Where (is it); and Why (is it there). Of course geographic pathology includes aspects of epidemiology since it is also concerned with prevalence and incidence and the interplay among complex causal factors, but it goes beyond epidemiology in its concern for the pathogenesis and the pathological effects of the disease under study.

One particular interest has been with infectious diseases as will be evident from my illustrations, but I emphasize that the MOD system is applicable to any study of diseases in which geographic/environmental factors are important. To be more specific, the computerized MOD project has two principal objectives. First, and most important, to develop a system which will include:

1. Standardized procedures for preprocessing medical information so that it is suitable for computer manipulation.

2. A storage/retrieval mechanism to act upon such preprocessed medical information, together with a complex editing program that will allow updating, will provide for immediate identification of material in conflict, will identify data sources, etc.

3. Programs whereby the computer can "instruct" a plotter to prepare contour maps reflecting quantitative aspects of incidence and/or prevalence of specific diseases, distribution of such causally related factors as animal reservoirs, insect vectors, climatic factors, and soil factors.

4. Programs whereby unmappable supporting information (to accompany the maps) can be printed out, thus extending the usefulness of the mapped medical information by pointing out certain limitations.

5. Programs allowing computer manipulation of the data to show significant interrelationships directly.

6. Programs whereby other types of graphic display of information can be generated to show direct cause and effect relationships (e g., line graphs) pertaining to prevalence and/or incidence of a given disease.

The second objective is to produce meaningful maps† (and other graphic displays) that show the distribution of a disease(s) in terms of prevalence, incidence, severity, etc., along with distributions of (and interrelationships among) selected causally related factors. Quantitative as well as qualitative aspects will be considered, with major emphasis on contour-type maps, the contour lines representing isarithms (isopleths).

The MOD project is primarily concerned with problems of data processing; there are many conventional aspects to these problems, but there are two unique aspects: (1) preprocessing a wide variety of physical and biomedical data, converting them to a compatible form which will allow subsequent computer manipulation and output of meaningful information relating to disease prevalence/incidence/ecological factors; and (2) developing a computer program to contour-plot disease data which are often represented by relatively sparse data points.

In relation to the first aspect, one of the major problems is to structure a data analysis vocabulary and develop a hierarchial system for the qualitative and quantitative characterization of disease and ecological information. This requires cutting across disciplinary boundaries, identifying the common denominator of the various jargons, and converting the narrative and tabular data into a miscible form.

Although the primary concern of this monograph is the processing of data, our work on the MOD project is such that I feel compelled to mention important limitations in the data available to us. There are many places in the world where the data base, dealing with many disease situations, is altogether inadequate for any meaningful collation, much less effective computer manipulation. We fully recognize that no system of information processing can convert bad data into good data!

†Technical Considerations: The various maps (graphic representations of spatially distributed data) will, ordinarily, be printed on transparent stock and overlayed onto base maps which contain physical and political geographic data, population data, etc. Of course, multiple transparent overlays, each presenting a different kind of information, may be layered togeth r to show the extent of pattern match, etc. I have already mentioned that quantitative aspects of the data can be handled by the use of isarithms.

However, there are large pools of data (derived from cultural anthropologists, economists, geologists, meteorologists, agricultural experts, epidemiologists, pathologists, etc.) relating to disease/environment situations which could be meaningfully collated and effectively computer-manipulated.

Many of the most important problems are characterized by the least minimally adequate relevant information, but we must identify what information there is and learn its limitations. We must work toward correcting deficiencies in the data base, but even more important, we must develop better methods of using what information is available.

We are at the stage of world development where many important judgments must be made in the absence of hard data. If we don't use what data is available what shall we use?

Now to return to the problem of data manipulation. An article in *Nature* describes the present situation well in terms of needs and accomplishments in the field of automatic data processing and computer mapping; these comments are very pertinent to the MOD project. It is stated that ". . . only a tiny proportion of the mass of demographic and climatic information collected by governments ever sees print in map form. Information is simply tabulated by area, and the possibility of spotting regularities or correlations—say the incidence of pellagra, family expenditure on food, and the provision of medical services in the eastern United States—is very remote indeed. Such interesting relations as have been found are the result of years of searching, and merely increase the sense of frustration that there is no better way of doing it."[1]

Because the principal output of our system is distribution maps, I will present a series of these maps to illustrate certain inherent problems, ways of attacking these problems, and some of our progress in overcoming these problems. I must emphasize that the maps which we have produced reflect efforts to develop technical methods, not—at this time—efforts to display new medical information. As a matter of fact, a major part of our work on computer production of maps has been based on a single convenient tabular source of data which was assembled some years ago by Dr. Jaques M. May (see TABLE 1). I will discuss our approach to structuring a data-analysis vocabulary which allows us to convert words of various qualitative and quantitative value to dots, or lines, or shaded areas on a map. But before getting into data structuring I will give a brief synopsis of our system design for handling input data (see FIGURE 1). I shall not pursue this aspect of the project further because I want to concentrate on data structuring and map production.

Our approach to a data-structuring vocabularly is shown in FIGURE 2, which illustrates the characterization of and relationships among

1. Low-Order Factors (LOF's) represent the most specific name or descriptor of a particular disease/environmental situation, e.g., pine trees, raccoons, degrees (of temperature), schistosomiasis, or Nigerian.

TABLE 1*

SOUTH AMEPICAN DATA USED IN MAP AND GRAPH GENERATION

Human Infection Rates of Schistosomiasis (*S. mansoni*)
Grouped by Province

Extracted from Source Document		Added by MOD Personnel	
Country	Inf. Rate (%)	LO, LA (nearest degree)	Inf. Rate (%)
Venezuela			
Aragua	24.8	−67, +9	24
Carabobo	9.9	−68, +10	9
Miranda	10.3	−66, +10	11
Distrito Federal	31.6	−67, +11	32
Guarico (Intradermo)	30	−66, +8	31
Dutch Guiana	Present	−55, +5	1
Brazil			
Maranhao	0.46	−45, −5	1
Piaui	0.04	−42, −8	1
Ceara	0.94	−39, −5	1
Rio Grande do Norte	2.32	−37, −5	2
Paraiba	7.53	−37, −7	8
Pernambuco	25.17	−38, −8	26
Alagoas	20.48	−37, −9	21
Sergipe	30.13	−38, −11	31
Bahia	16.55	−42, −12	17
Espirito Santo	1.63	−41, - 20	2
Minas Gerais	4.41	−45, −18	4
Rio de Janeiro	0.10	−43, −22	1
Parana	0.12	−49, −22	1
São Paulo	Present	−52, −24	1
Santa Catarina	0.00	−51, −27	0
Goias	0.03	−49, −13	1
Matto Grosso	0.007	−55, −17	1

*From May, J. M., Ed. 1961. Studies in Disease Ecology. Hafner Publishing Co., Inc. New York, N. Y.

2. Middle-Order Factors (MOF's) represent the set of all LOF's which describe the same aspect of disease/environmental situations. A MOF represents a category of descriptive terms that pertains to only one particular aspect of disease—it requires LOF's to give it substance. Examples are given in the following list.

MOF	LOF's Making Up the MOF
Measure	Occurrence, abundance, point prevalence, period prevalence, incidence, inches
General kind of disease	leptospirosis, schistosomiasis
Specific disease agent	*Leptospira pomona*, *L. canicola*, *Schistosoma mansoni*, *S. japonicum*

There are two kinds of MOF's: C-MOF's (Common-MOF's) and O-MOF's (Optional-MOF's). C-MOF's are those MOF's which, in the MOD system, are common to or will accompany (as necessary descriptive elements) every bit of mappable data. C-MOF's include these, and only these, MOF's:

C-MOF	LOF'S Making Up the C-MOF
1. Security classification of the data	Top secret, secret, confidential, restricted for official scientific use only, unclassified
2. Primary source document identification. (The primary source document is the paper which originally reported the data.)	Abbreviated bibliographic citation, i.e., author(s), date, journal, book, volume, and page. (This MOF will always be used, whether the data comes from primary or from secondary source documents.)
3. Secondary source document identification. (The secondary source document is a paper which references or quotes data already reported.)	Abbreviated bibliographic citation, i.e., author(s), date, journal/book, volume, and page. (If the data being structured is extracted from its primary source document, this MOF will be left blank or not used.)
4. Professional evaluation of data source (i.e., author, organization or institution, or source document)	More reliable, less reliable, reliability not assessed.
5. Computer evaluation of data point (to be calculated internally by the computer)	a number
6. Time period for which the data applies	1963, 1960–1964, June 1965, pre-1966

O-MOF's, on the other hand, are those MOF's which will not necessarily accompany every bit of mappable data. At least one O-MOF will be part of every bit of mappable data; however, both the specific O-MOF's used and their number can vary widely.

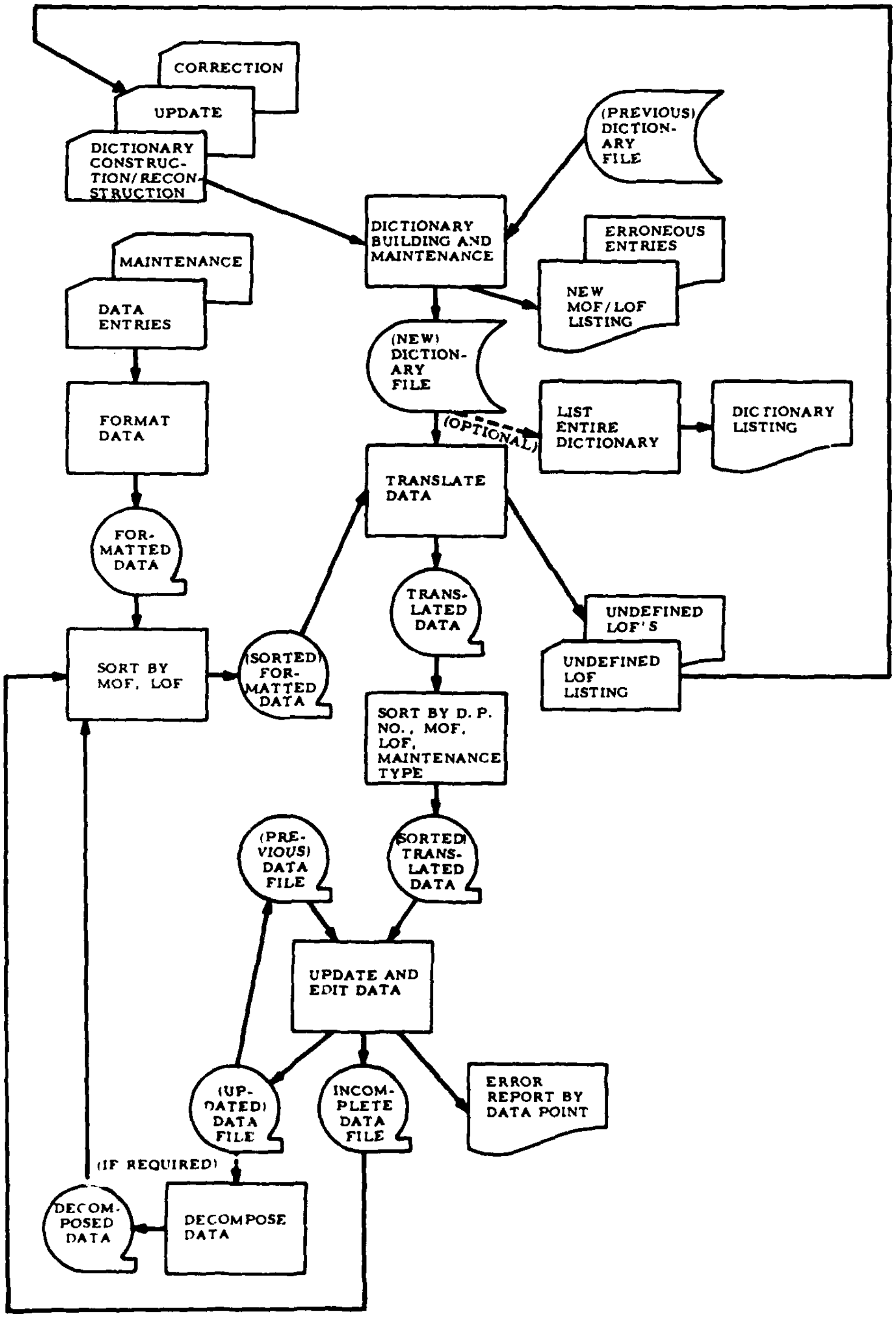


FIGURE 1. System design for handling input data.

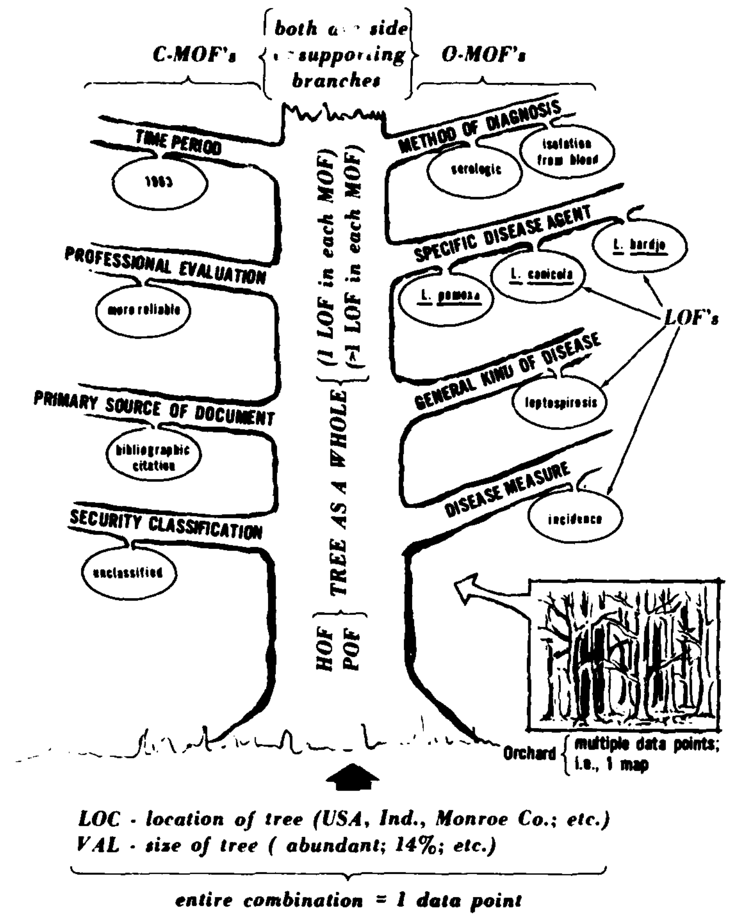


FIGURE 2. A tree-like (hierarchal) relationship of terms for disease/environmental descriptions. See text for explanations.

High-Order Factors (HOF's) represent a specific combination of LOF's, in which each LOF belongs to (is drawn from) a different MOF, i.e., in which no MOF contributes more than one LOF.

Poly-Order Factors (POF's) represent a specific combination of LOF's, in which at least two LOF's belong to (are drawn from) the same MOF, i.e., in which at least one MOF contributes more than one LOF. For example, the following is a HOF: Incidence (measure) of schistosomiasis (general kind of disease) due to *Schistosoma mansoni* (specific disease agent) in Indians (animal host infected) as determined by fecal analysis, during 1958 (time period for which data applies). This statement would be a POF if, at specific disease agent an additional LOF was added (from the same MOF), e.g., and *Schistosoma hematobium* or if, at animal host infected an additional LOF (from the same MOF) was added, e.g., and Caucasians. FIGURE 3 should help to clarify this relationship.

Factor is a general term including LOF's, MOF's, HOF's, and/or POF's. LOF's, MOF's, HOF's, and POF's can be viewed together as a kind of hierarchy or a kind of matrix, as shown in FIGURES 2 and 3.

In order to map factors, location and values must be coupled with them. Location (LOC) is an exact geographic position, stated as precisely as possible, of each bit of mappable data. For purposes of the MOD system, each bit of data can have two, and only two, LOC's accompanying it:

1. Geographic location by longitude and latitude	W 088°31′, N 37°29′; W 044°18′, S 17°09′
2. Geographic location by political unit	Pope County, Ill. USA, North America Minas Gerais province Brazil, South America

Value (VAL) is an alphabetic and/or numerical symbol expressing one member of the set of all possible results, that result-set describing the functional relationship governing a specific factor (HOF or POF) as it ranges over the set of all possible LOC's. For example: 0, 1, 2, 3, . . . : 0, 0.01, 0.02, 0.07, . . . ;

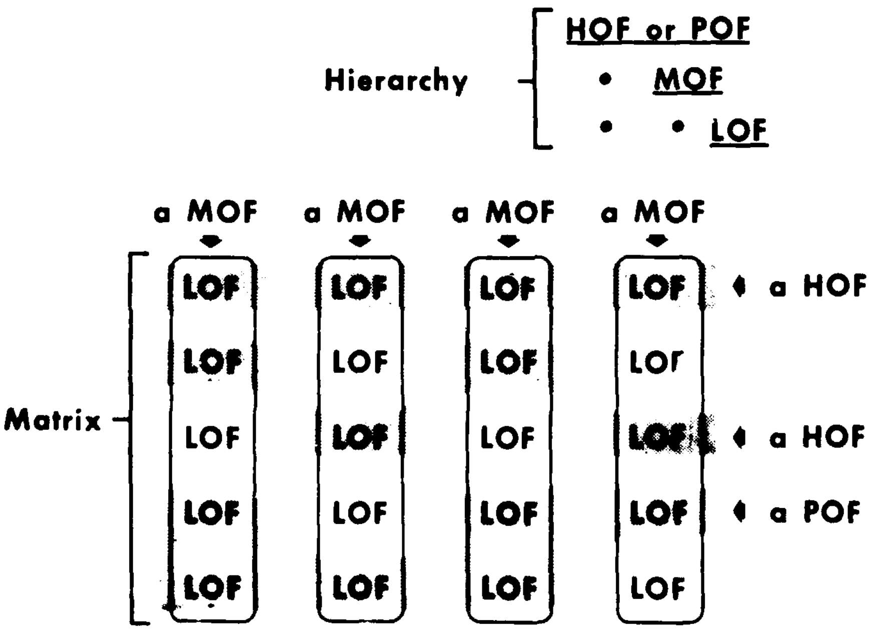


FIGURE 3. Relationship among term categories. See text for explanations.

0–10, 10–20, 20–30; . . . ; absent, present; absent, rare, common, abundant; shale, limestone, sandstone, granite.

The LOC describes where a disease/environmental situation was studied. The factor (HOF or POF) specifies what aspect of the disease/environmental

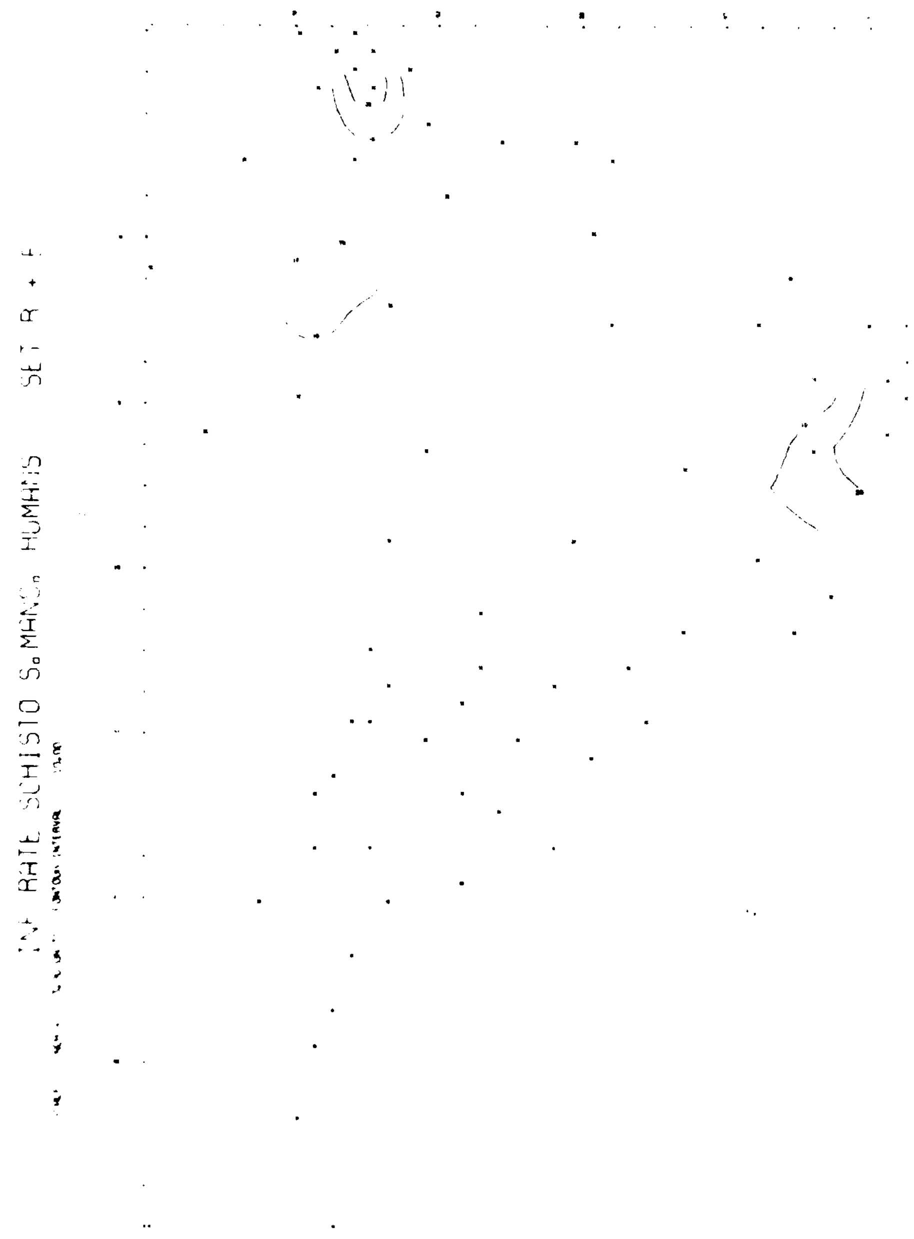


FIGURE 4

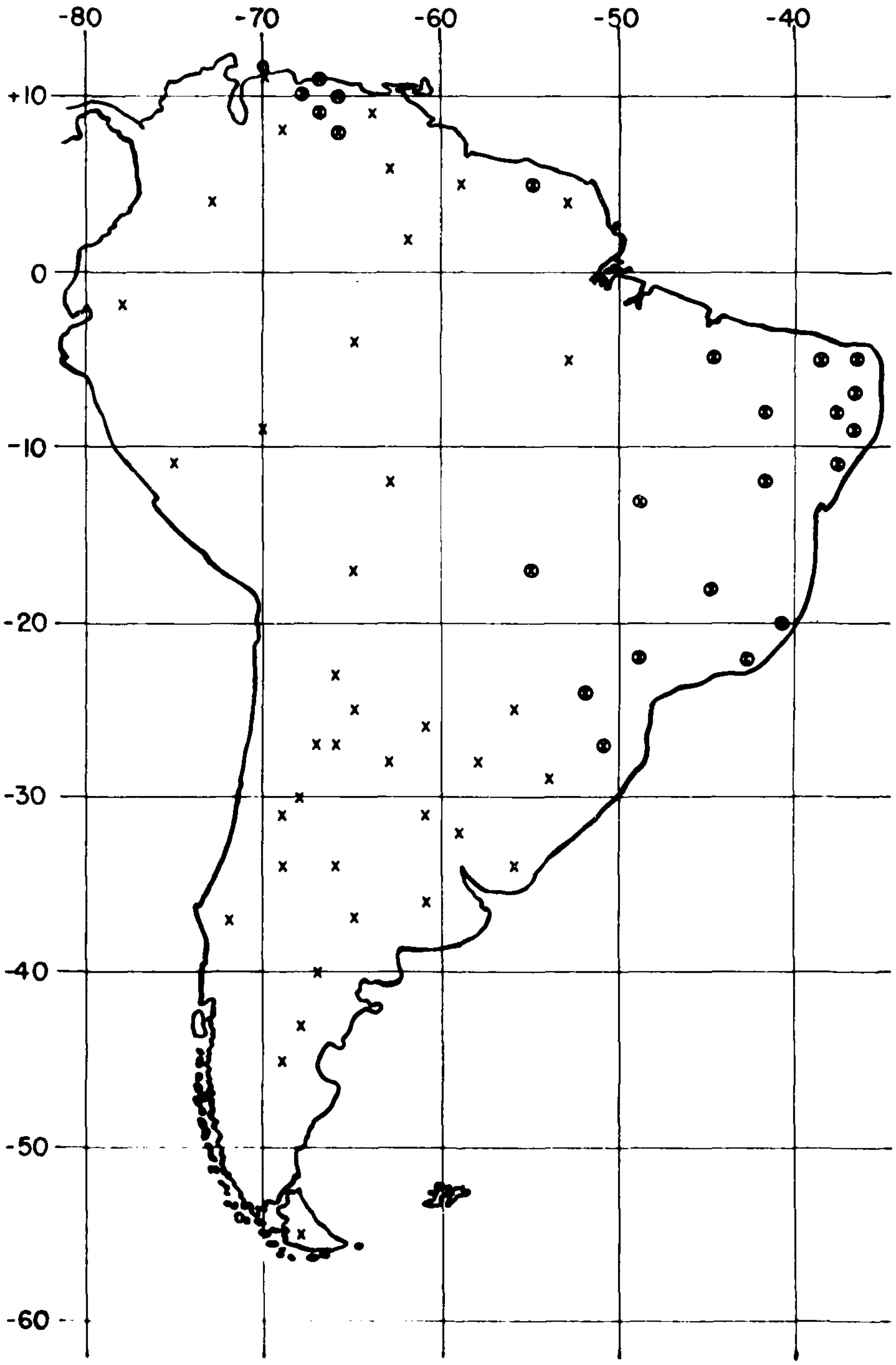


FIGURE 5.

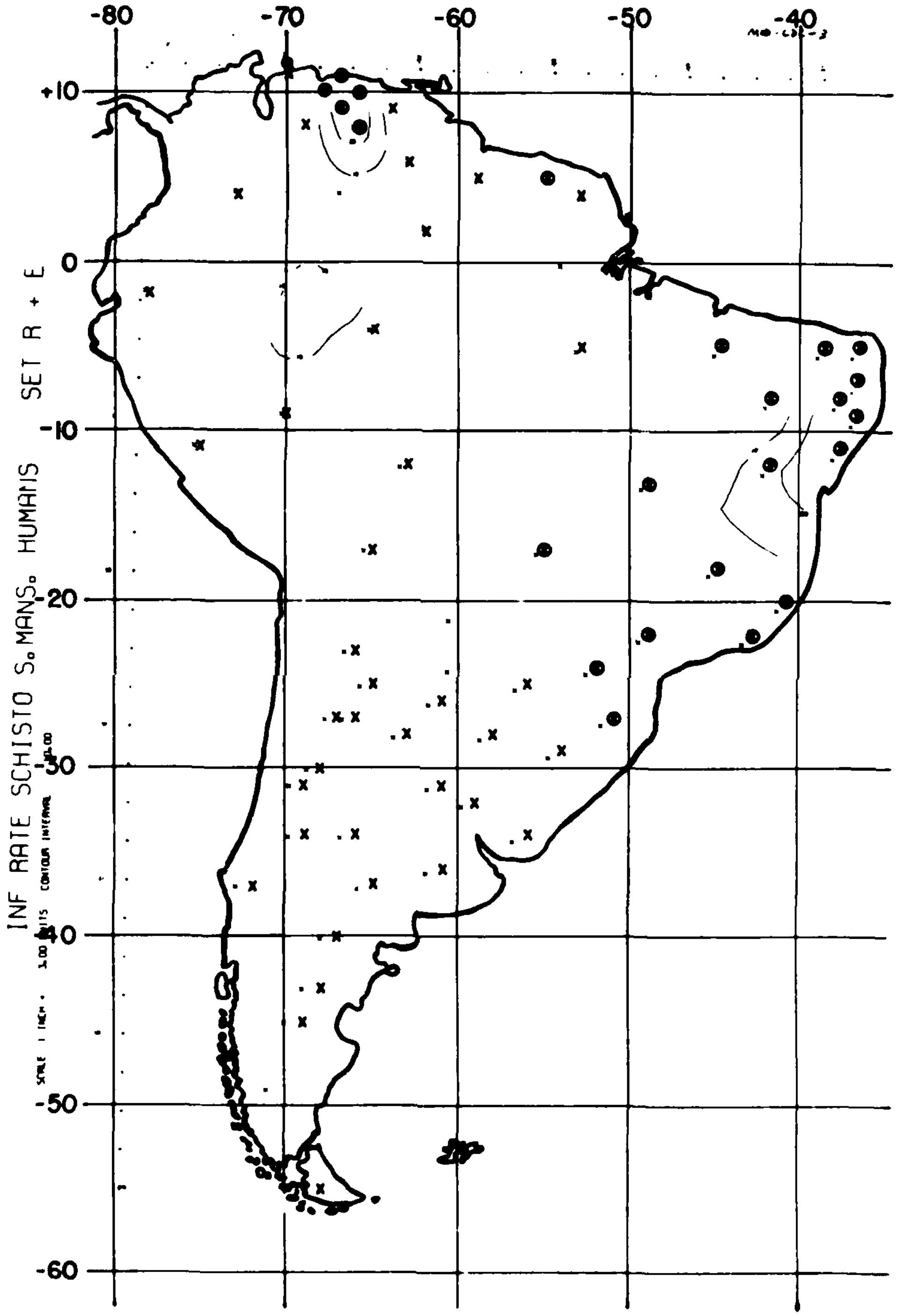

-80
-70
-60
-50
-40
+10
0
-10
-20
-30
-40
-50
-60
INF RATE SCHISTO S. MANS. HUMANS SET R + E
SCALE 1 INCH = 1.00 UNITS CONTOUR INTERVAL 10.00

FIGURE 6.

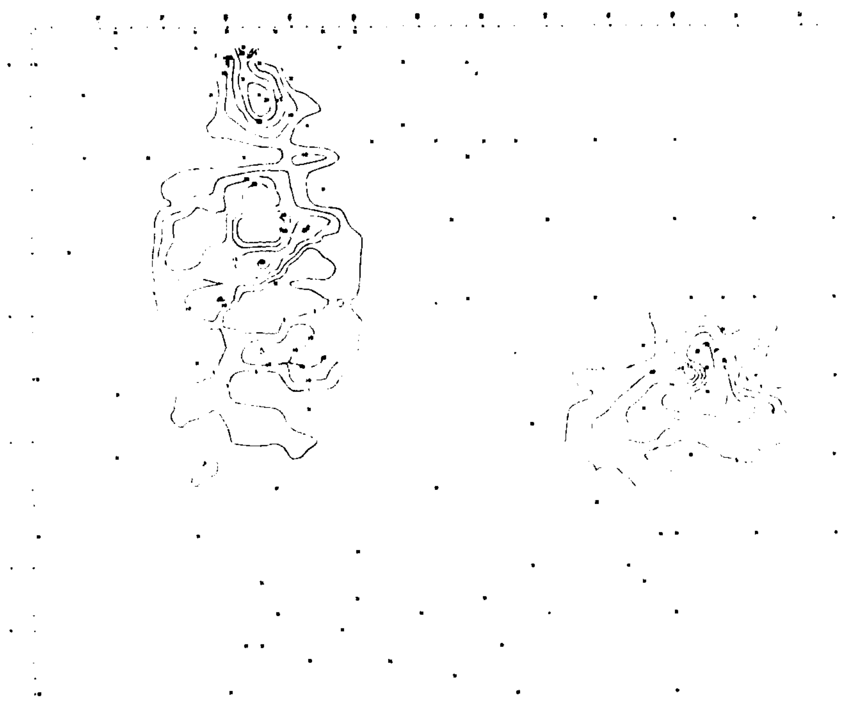


FIGURE 7.

situation was studied. The VAL describes the result or conclusion reached by the studies. A data point is the combination of a specific LOC, a specific HOF or POF, and a specific VAL. In general, LOF's and MOF's cannot be mapped, because by themselves they do not convey enough information to be meaningfully mapped. However, HOF's and POF's can be meaningfully mapped, with each HOF or POF serving as the description (legend) of its corresponding map.

Turning now to the production of maps, contour-type maps have proved to be extremely useful tools in studying various environmental characteristics or factors, particularly those considered by the earth sciences. Such maps have great potential value in many studies of disease situations. Thus, a general discussion of this type of map is altogether pertinent to the present phase of the MOD project.

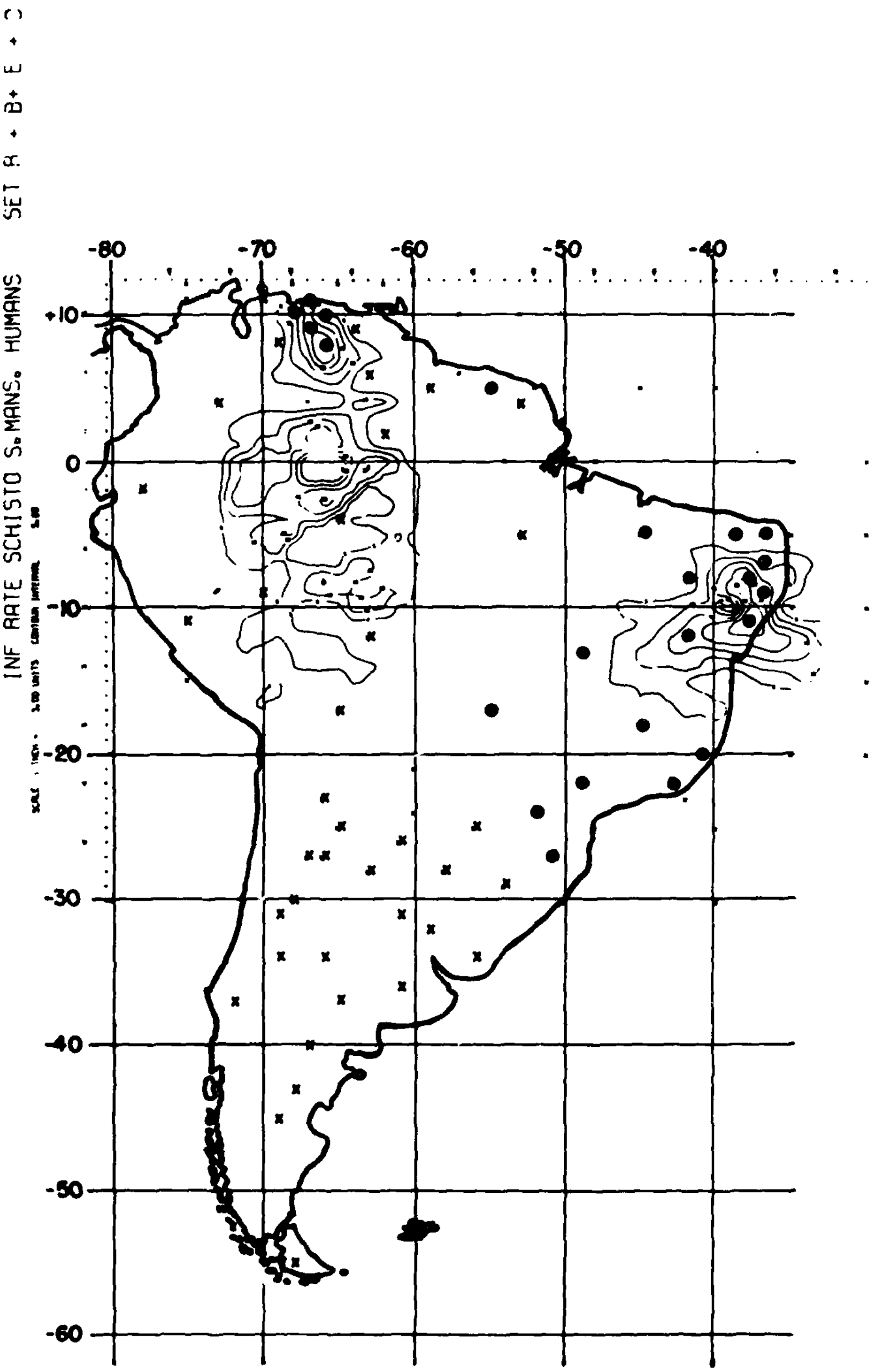

INF RATE SCHISTO S. MANS. HUMANS SET A + B+ E + C
SCALE 1 INCH = 5.00 UNITS CONTOUR INTERVAL 5.00
-80
-70
-60
-50
-40
+10
0
-10
-20
-30
-40
-50
-60

FIGURE 8.

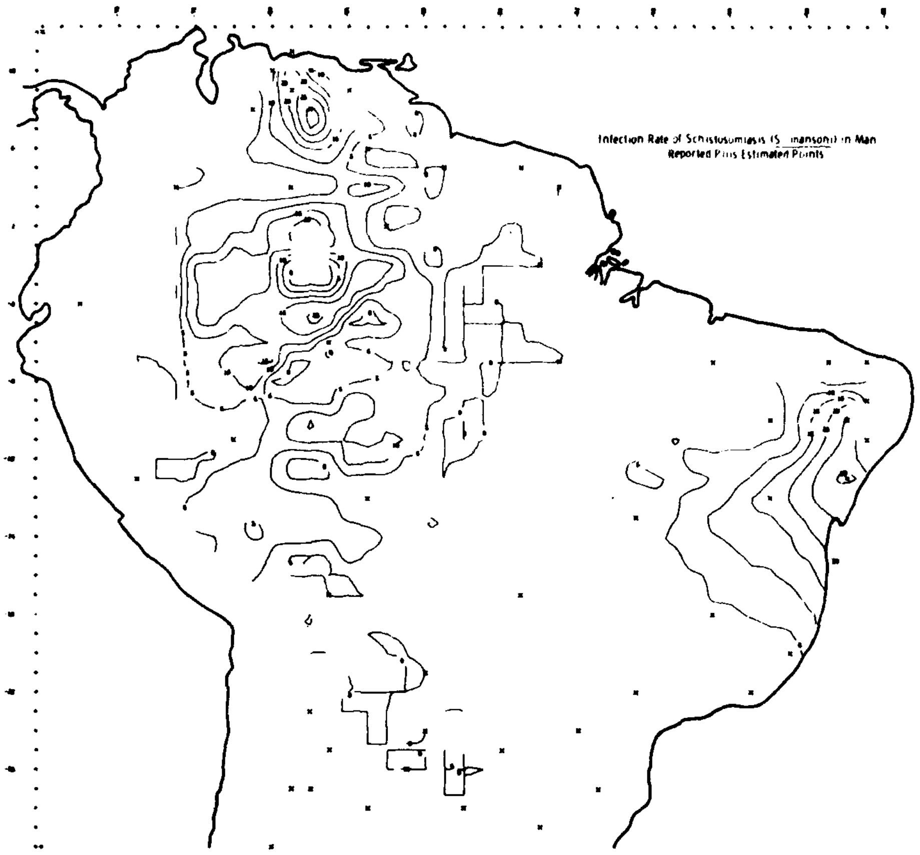

Infection Rate of Schistosomiasis (S. mansoni) in Man
Reported Plus Estimated Points

FIGURE 9.

In essence, a contour-type map is a device by means of which a three-dimensional, complex, geometrical figure can be represented on a two-dimensional plane surface. The map accomplishes this by a set of form lines, the contours (sometimes called isolines or isarithms), which outline according to well-defined rules the shape of that complex geometrical figure.

Because the geometrical figure being represented by the map is three-dimensional, this figure can be treated as a set consisting of many ordered triplets of numbers (X_i, Y_i, Z_i). For each specific pair (X_n, Y_n), one and only one Z value, Z_n, exists, i.e., $Z = F(X, Y)$. Theoretically, the variables X and Y can represent values of any conceivable independent disease/environmental factors; if this is allowed, the result is a graph showing the relationship among three disease/environmental factors. In order to make a geographic map of a particular disease/environmental factor, the X value is taken to be the longitude (LO) of a geographic point locality; the Y value is the latitude (LA) of the locality. The Z value is the VAL of a specific

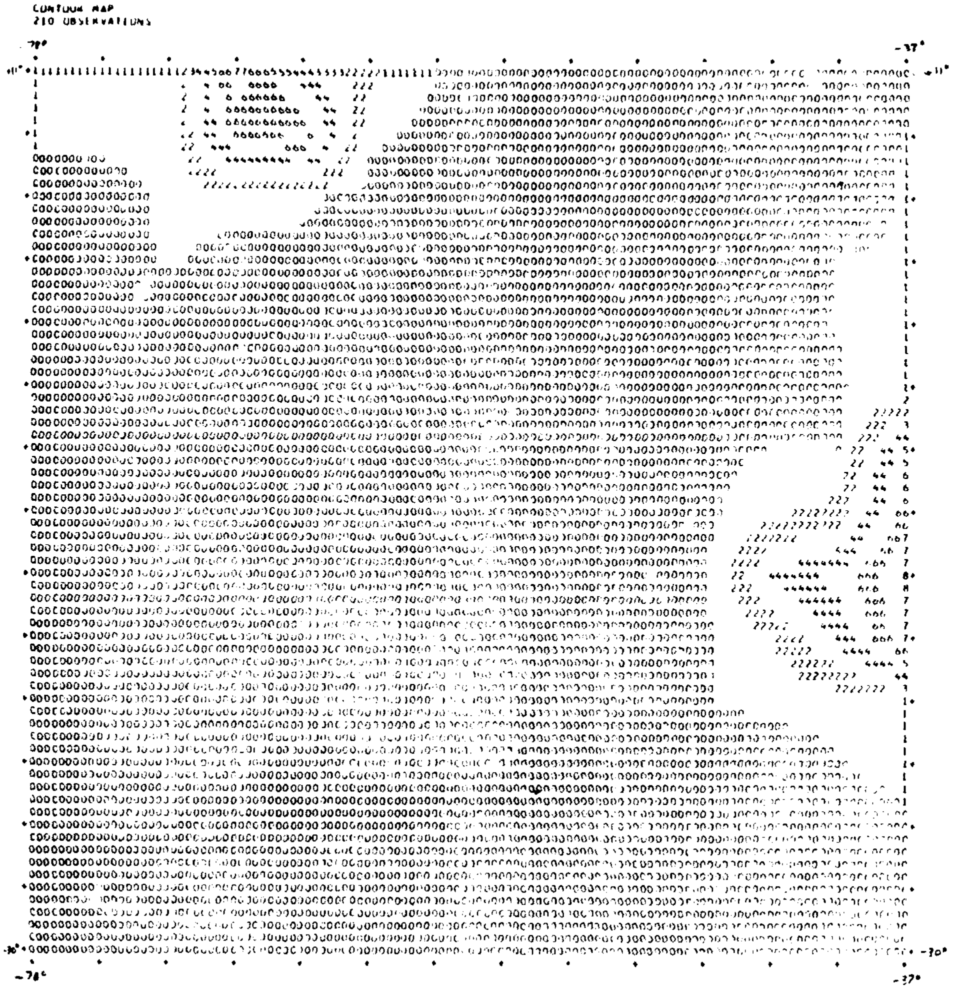


FIGURE 10.

factor at that particular geographic point locality. Thus, in making a map, the (X, Y, Z) triplets become special cases, i.e., (LO, LA, VAL) triplets. Each contour is a line that connects an infinite number of contiguous/adjacent geographic point localities (LO, LA) which have the same VAL for the factor being mapped.

I will briefly discuss plotter programs in connection with specific maps. FIGURE 4 shows one of our very first attempts to plot schistosomiasis infection rate data (from TABLE 1) using a CDC program and their CDC-3600 computer, and a 30 inch drum-type Calcomp plotter, operating off-line. These isarithms were constructed from the data points located as shown in FIGURE 5. When the lines of FIGURE 4 (on transparent acetate) were laid over the base map (FIGURE 5), the result was as shown in FIGURE 6

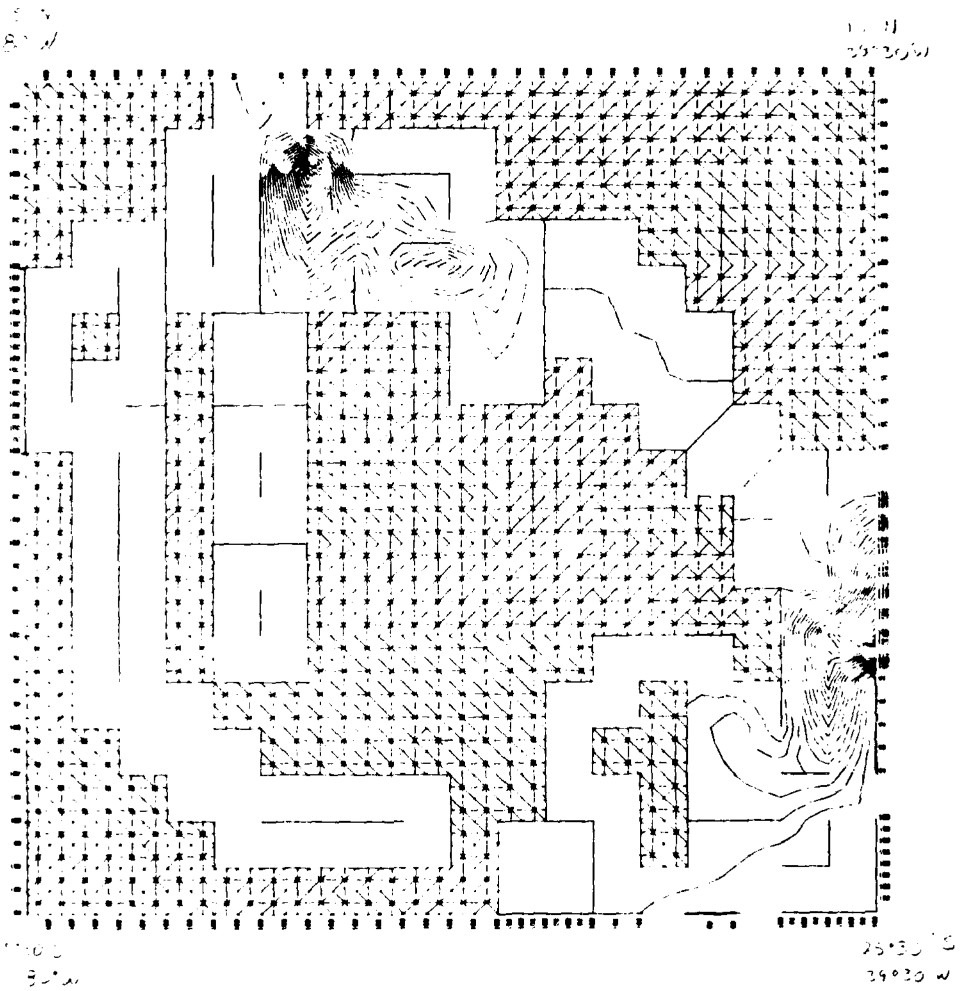


FIGURE 11.

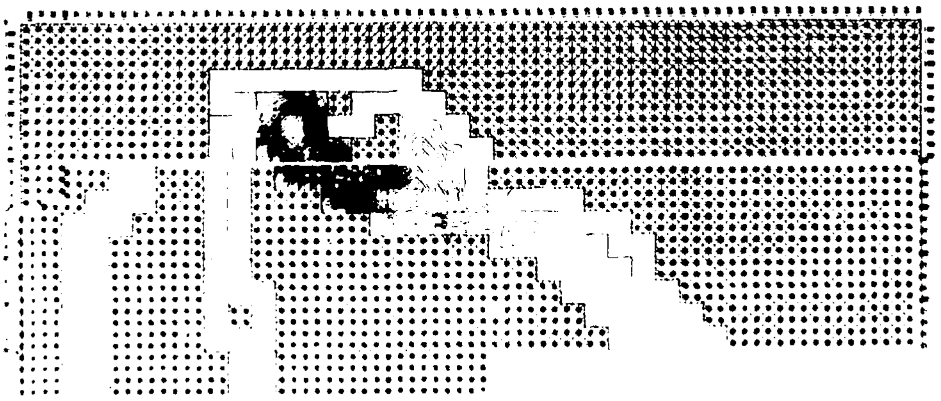
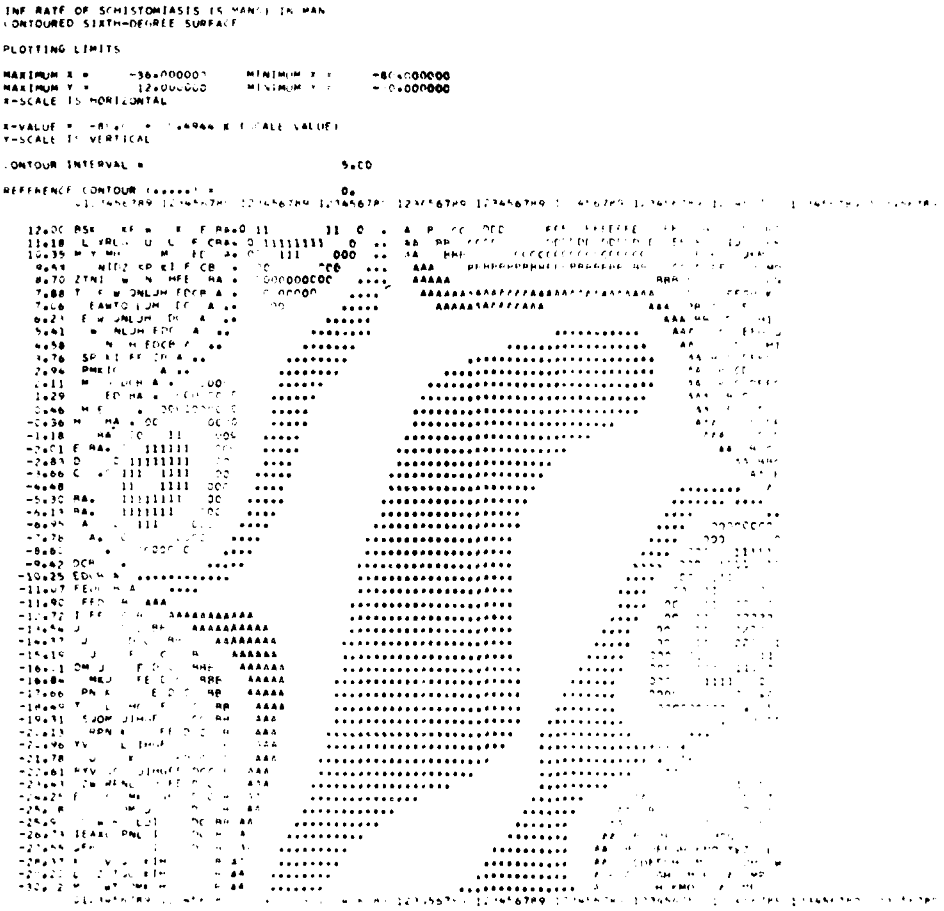


FIGURE 12.

Changing grid sizes, modifying queries, etc. led to a long succession of maps, most better, a few worse. Several of these are shown: FIGURES 7 and 8 (composite—acetate overlay and base map) indicate a schistosomiasis infected population living out in the ocean (!) because of an unrealistic extrapolation between positive value data points and (inserted) 0 value data points. This situation is corrected in FIGURE 9, but spurious data is represented in western Brazil and northern Argentina because of fruitless efforts to contour-relate sparse low incidence values.



```
INF RATE OF SCHISTOMIASIS (S MANSONI) IN MAN
CONTOURED SIXTH-DEGREE SURFACE

PLOTTING LIMITS

MAXIMUM X =      -36.000000     MINIMUM X =      -80.000000
MAXIMUM Y =       12.000000     MINIMUM Y =      -30.000000
X-SCALE IS HORIZONTAL

X-VALUE =  -80.0  +  0.4944 X (SCALE VALUE)
Y-SCALE IS VERTICAL

CONTOUR INTERVAL =                                5.00

REFERENCE CONTOUR (......) =                      0.
```

FIGURE 13.

FIGURE 10 represents an effort to get fast read-out. This seems a plausible "quick look" method (based on data in TABLE 1) to determine whether or not the higher resolution, but slower and more laborious, plotter technique would be justified.

FIGURES 11 and 12 (based on data from TABLE 1) were produced using a Naval oceanographic program. They illustrate the unrealities of data plots if one bases fine analyses on relatively sparse data, somewhat analogous to averaging "rough" Figures such as 10, 9, 7, 3, and carrying out the result to three decimal places.

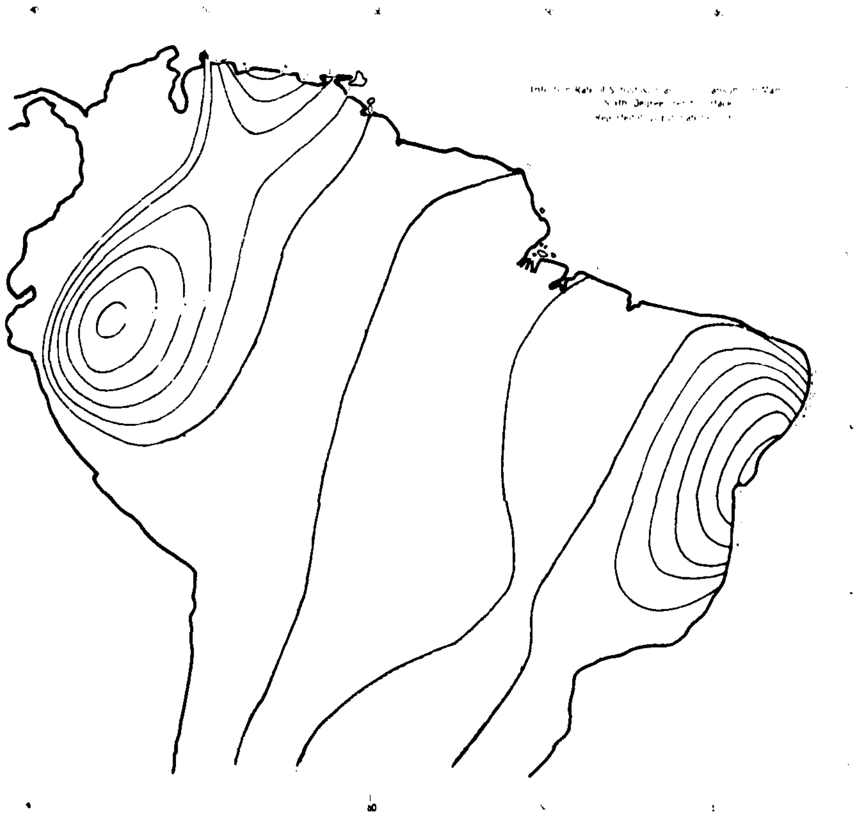


FIGURE 14.

FIGURE 13 presents a different approach to extrapolation in that it is based on a sixth-degree trend surface computation.

FIGURE 14 is a hand-drawn map based upon the computer produced lines of FIGURE 13. For reasons not apparent to us the points of highest disease incidence have "drifted."

The most promising system for handling relatively sparse data points is a method described by Tobler[2] using a simpler algorithm. In essence, it is to use only the three closest points surrounding a grid point and to fit a plane through

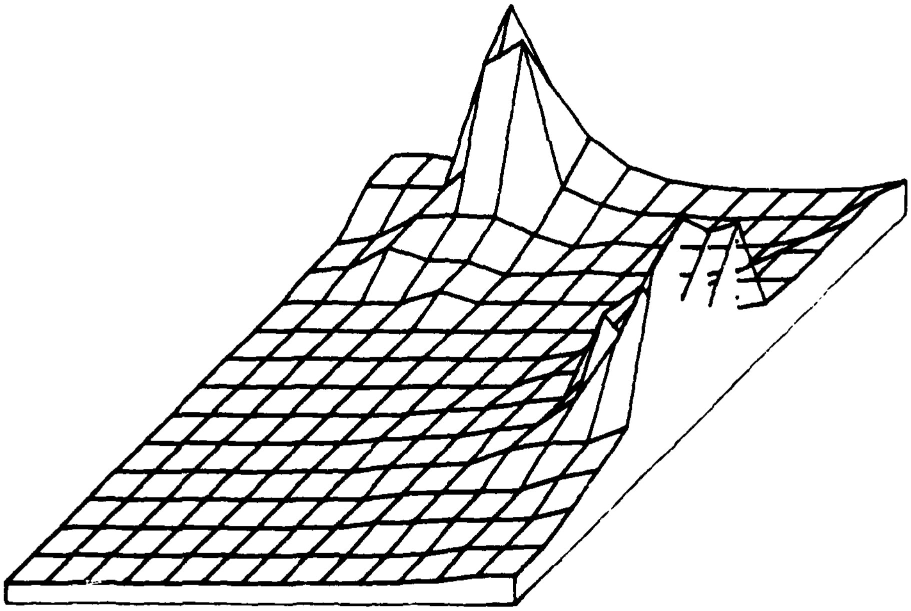


FIGURE 15.

them. The grid point, therefore, lies on this plane and its value may be computed. The method, as it applies to this problem, is as follows:

1. Compute the distance from the grid point to all observed points (which are assumed to be randomly distributed).
2. Out of all of these points, find the nearest three which surround the grid point in question. Fit a plane, Z = AX + BY + C, through the three points by solving the system of simultaneous linear equations necessary to fit a plane through points whose X, Y, and Z coordinates are known.
3. Calculate the estimated value at the desired grid point by inserting its coordinates into the equation.
4. Continue to the next grid point in the row. Stop when all rows have been examined.

This method has been simulated manually by us and the validity of the concept demonstrated as illustrated in FIGURES 15 and 16.

Potential Applications of the MOD system, in addition to producing distribution maps per se (and other graphic displays) and in helping to determine causal relationships include: (1) use in predicting the probability of changes in incidence/character of specific diseases as a consequence of particular changes in ecology, and (2) use in developing mathematical models by which one may predict major changes in disease incidence, e.g., epidemics.

Summary

The MOD$_{\wedge}$ Project is an effort to: (1) characterize input data (relating to disease/environment) in such a way that they can be stored and readily retrieved in context by a computerized system which, (2) using these data can meaningfully relate the prevalence, incidence, and character of a disease to a variety of direct and indirect causal factors, always with a time and location characteristic, and (3) output the information directly in map form.

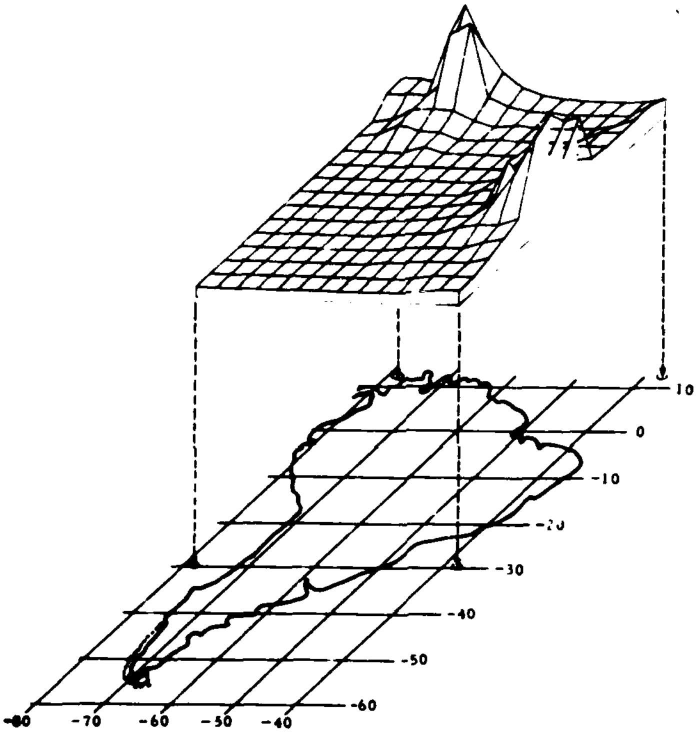


FIGURE 16.

Acknowledgments

I would like to express my appreciation to my associates, Roger J. Cuffey, William H. Watson, Jr., Jerome Morenoff, Wayne L. Richmond, and Joseph D. H. Sidley.

References

1. MAPS BY MACHINE. 1967. (Editorial). Nature (London) 213: 1166-1167.
2. TOBLER, W. R. 1965. Automation in the preparation of thematic maps. Cartographic J. 2(1): 32-38.